DOCUMENT RESUME

ED 206 656 TM 810 548

AUTHOR Tobin, Kenneth G.; Capie, William
TITLE An Empirical Investigation of the Stability of Variance Components and Dependability Coefficients Derived from Teacher Performance Data.
PUB DATE Apr 81
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; Competence; *Teacher Evaluation; Teaching Skills
IDENTIFIERS *Generalizability Theory; Sample Size; *Teacher Performance Assessment Instruments

ABSTRACT

Teacher performance data was used in this investigation to determine whether the stability of variance components and generalizability coefficients could be obtained by increasing sample sizes. A section of the Teacher Performance Assessment Instruments (TPAI) was utilized to assess 2 competencies for each of 750 teachers Each competency, defined by five performance indicators rated on a five point scale, was evaluated twice. Three sets of analyses were performed for each competency: the number of teachers was increased in increments of 50-700: 5 replications based on data from 50 teachers and: 5 replications based on data from 150 teachers. Variance components and frequency of negative variance components were estimated for each analysis and generalizability coefficients calculated for each teaching competency. Results indicated that stability increased as the sample size increased. A minimum of 150 teachers appears to be required in order to obtain stable variance components and generalizability coefficients and to maintain the number of negative variance components. (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206656

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AN EMPIRICAL INVESTIGATION OF THE STABILITY OF VARIANCE COMPONENTS AND DEPENDABILITY COEFFICIENTS DERIVED FROM TEACHER PERFORMANCE DATA

Kenneth G. Tobin
Center for Science Education
Mount Lawley College
Australia 6050

William Capie
Department of Science Education
University of Georgia
Athens GA 30602

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
K.G. Tobin
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 548

A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April, 1981

AN EMPIRICAL INVESTIGATION OF THE STABILITY OF VARIANCE COMPONENTS AND DEPENDABILITY COEFFICIENTS DERIVED FROM TEACHER PERFORMANCE DATA

With the advent of competency-based teacher certification, emphasis has been directed toward reliable assessment of teacher performance. Because of the multiple variance sources extant in teacher performance measures, the application of generalizability theory for norm-referenced and criterion-referenced decisions has assumed considerable importance. For example, teacher performance measures collected in field trials of the Teacher Performance Assessment Instruments (TPAI) (Capie, Johnson, Anderson, Ellett, and Okey, 1979) have been used as a basis for estimating generalizability coefficients and Brennan's $\emptyset$ (λ) (Brennan, 1978). An assumption underlying the use of the measures for this purpose is that the estimated variance components are representative of population values. Recently the validity of this assumption has been questioned by researchers such as Smith (1978, 1980).

Little is known about contextual factors that influence stability or sampling characteristics of variance component estimates. Using simulated data, Smith (1978) found considerable dispersion of variance component estimates when the number of levels for each facet in a generalizability study (G-study) was less than ten. As a consequence,

Smith advocated the use of G-study designs which incorporated more than ten levels for each facet as well as a large sample size. There is some doubt, however, that results involving simulated data or pencil and paper measures are applicable to other types of data. Also, practical concerns often preclude the use of ten levels of each facet. For example, answers to be obtained from a G-study may necessitate the use of an existing data collection system. This is the case in Georgia where the TPAI are to be used for teacher certification. Existing procedures require teachers to be assessed on two occasions by an external data collector, a school administrator, and a peer teacher. Since the data collection procedures have been determined, the major purpose of a G-study is to estimate generalizability coefficients or Brennan's $\emptyset$ (λ) so that confidence intervals can be calculated for measures of teacher performance and established performance criteria respectively. Consequently, a strong case exists for utilizing the same design for collecting G-study and D-study data. Another compelling argument is that the number of levels for fixed facets should not be increased above the fixed level. The TPAI data collection system regards two facets as fixed : the number of performance indicators that define each teaching competency; and the occasion of measurement. In each case the number of levels involved is much less than ten. As a consequence, an investigation was planned to determine whether tolerable variance component variability could be obtained by increasing sample size.

According to Smith (1978) the variance of an estimate of a variance component takes the form:

$$\text{Var}\ (\hat{\sigma}^2) = \frac{2\sigma^4}{\mu}$$

where $\hat{\sigma}^2$ is the variance component estimate; $\hat{\sigma}^4$ is the squared variance component; and u represents the degrees of freedom for the variance. Thus, for all variance components involving the teacher, the number of teachers in the study is represented in the denominator of the above expression in the degrees of freedom term. Unfortunately the algebraic expressions that represent the variance of most variance component estimates are quite complex for four-facet designs. However, the expressions for the three-way interactions are comparatively simple and will be used for illustrative purposes. For example, the variance for the teacher by indicator by occasion interaction is described by the expression presented below:

$$\text{Var}\ (\hat{\sigma}^2_{\text{TIO}}) = \frac{2}{(T-1)(I-1)(O-1)} \{\hat{\sigma}^4_{\text{TIO}} + \frac{\hat{\sigma}^4}{R^2}\}$$

where T is the number of teachers in the study; I is the number of performance indicators which define the competency; and O is the number of occasions on which teachers are assessed.

The expression shows that the magnitude of the variance is dependent on several factors. However, a low number of levels on any facet may be compensated by a large sample size. Thus, for norm-referenced applications of teacher performance assessment data, the main implication for stability of variance component estimates, is that a large sample of teachers should be included in the study. By

so doing, the dispersion of the variance components that define the generalizability coefficients can be reduced.

The primary concern of this study was to examine variability of variance component estimates as sample sizes were increased and to investigate variability from replicated designs. A secondary purpose of the study was to document the incidence of negative variance component estimates. Since there is a non-zero probability that an ANOVA estimate of a variance component will be negative, the effect of increasing sample size on the incidence of negative variance component estimates was of interest.

Data Sources

The TPAI assess fourteen competencies related to planning and classroom performance. In this study separate analyses were conducted for two competencies each defined by five performance indicators rated on a five point scale. The first competency consists of indicators designed to assess ability to plan instruction to achieve selected objectives. The information required to rate the indicators is obtained from a portfolio of lesson plans or from a teacher interview. In contrast, the data for assessing the second competency are obtained from direct observation of one class period. A sample performance indicator for the second competency is contained in Figure 1.

Insert Figure 1 about here

Each teacher was assessed on two occasions by a colleague, an administrator, and a data collector external to the school system. A sample data set was selected for each competency from 1980 performance data for all teachers in Georgia with zero or one year of teaching experience. Only teachers with complete data were included in this sample, herafter referred to as the selected sample. The selected sample consisted of approximately 750 teachers.

Procedures

Three separate sets of analyses were conducted for each competency: the number of teachers was increased in increments of 50 to a sample size of 700; five replications based on performance data from 50 teachers; five replications based on performance data from 150 teachers. Replication samples were drawn, without replacement, from the selected sample for each competency. For each analysis fifteen variance components were estimated from expected mean square terms obtained from ANOVA. The frequency of negative variance component estimates was also determined for each analysis.

Generalizability coefficients were calculated for each teaching competency using a model in which each facet was regarded as random. Brennan's $\emptyset$ (λ), a coefficient that is applicable for criterion-referenced applications of the teacher performance measures, was also calculated for each competency.

Results

The results for each competency are presented separately in the sections that follow.

Plans Instruction to Achieve Selected Objectives.

The data in Table 1 show the variation that occurred in variance component estimates for a number of sources as sample size was increased from 250 to 700. Variation in the magnitude of the generalizability coefficient and Ø (λ) can also be observed.

Insert Table 1 about here

The number of negative variance components obtained was negligible for sources pertinent to norm-referenced applications involving differentiation of teacher performance. The only negative component obtained was for the teacher by indicator by rater interaction at a sample size of 50. When all variance components were considered, 12 percent were negative with most occurring for the rater by occasion effect and the indicator by rater by occasion effect.

The results in Table 2 indicate that the standard deviation for the variance components was smaller for estimates based on samples of 150 than for estimates based on samples of 50.

Insert Table 2 about here

Communicates with Learners

The data included in Table 3 show the variations that occurred in variance component estimates, the generalizability coefficient, and $\emptyset(\lambda)$ as sample sizes were increased from 200 to 700. The results in Table 4 provide a basis for comparing estimates based on sample sizes of 50 and 150 respectively.

Insert Table 3 and Table 4 about here

The number of negative variance component estimates was approximately eight percent. No negative values were obtained for variance components used to estimate the generalizability coefficient for differentiating teacher performance in norm-referenced situations.

Discussion

Although the two measures of teacher performance were based on different data sources, the results for each competency were similar with respect to:

1. the variation in variance component estimates;
2. the generalizability coefficient;
3. $\emptyset(\lambda)$ for replicated sample sizes of 50 and 150;
4. trends observed as sample size was increased to approximately 700 teachers;

5. the incidence of negative variance components for a range of effects.

Variation in the estimates of variance components, and dependability coefficients was smaller for sample sizes of 150 teachers than for sample sizes of 50 teachers. This outcome is consistent with the findings of Smith (1978) who concluded that the most promising solution for estimating variance components was to increase sample sizes. The coefficients of variation reported in Tables 2 and 4 suggest that variability associated with estimates based on samples of 50 teachers (1500 observations) is too large. For example, the coefficient of variation for the generalizability coefficient for competency one was 17 percent. By comparison, the coefficient of variation was less than ten percent for the generalizability coefficient and $\emptyset$ (λ) for each competency for sample sizes of 150 (4500 observations).

The results reported in Tables 1 and 3 show that variability for all estimates was small as sample size was increased from 200 to 750. Of particular interest was the stability of the generalizability coefficient and $\emptyset$ (λ) in this range. From a consumer perspective, however, the important question relates to the operational definition of tolerable variability. Any definition is likely to be context dependent, however, for teacher performance assessment measures a coefficient of variation of less than ten percent was regarded as a satisfactory criterion for stability. As a consequence, the estimates obtained for the samples of 150 teachers were regarded as

satisfactory. If the total number of observations is considered the crucial variable, estimates based on 4500 observations appear to have tolerable variability.

Smith (1978) also reported greater accuracy in variance component estimates when the algebraic expression for the expected mean square contained fewer terms. This trend was not strong in the results obtained from this study. This may be attributed to comparative instability of interaction terms involving occasions and raters which had two and three levels respectively.

The proportion of negative variance components obtained was approximately ten percent. Negative variance components were rarely obtained for an effect involving teachers. Since teacher performance data are usually used to differentiate teacher performance, negative components are of greatest concern when they occur for the teacher effect or for interaction terms involving teachers. When sample sizes of 50 were used, negative variance components were rarely obtained for effects invrlving teachers. The general trend was that the number of negative variance components decreased as the sample size was increased. The most common negative estimates were obtained for the rater by occasion effect and the indicator by rater by occasion effect. In each case the source of variation was associated with a comparatively low number of degrees of freedom.

The results of this study should be interpreted with caution since they are based on a relatively small number of samples. This is a

necessary consequence of using measures of teacher performance rather than simulated data. Although the results have greatest implications for D-study applications of the TPAI, the findings also provide general guidelines on development and research in teacher performance assessment. Comparatively stable variance component estimates may be obtained when selected facets have fewer than ten levels provided that large sample sizes are used. This finding has practical advantages for facets such as assessment occasion and rater type where increases in the number of levels substantially influence costs. For facets such as the number of performance indicators that define a competency there could be some advantage in incorporating ten performance indicators to increase the stability of G-study estimates.

Conclusion

Smith (1980, 1978) suggested that a minimum number of ten levels of each facet should be incorporated into a G-study design in order to assure stable variance component estimates. Whilst this may be a desirable goal, practical considerations will often preclude the use of ten levels. This study has shown that tolerable variation in variance components of interest, the generalizability coefficient for differentiating teachers, and $\Phi(\lambda)$ for criterion-referenced applications were obtained from samples of 150 teachers when facets had five, three, and two levels respectively. If the purpose of a G-study is to estimate a generalizability coefficent or $\Phi(\lambda)$ then suitable coefficient estimates might be obtained with approximately 4500 observations.

References

Brennan, R.L., Extensions of Generalizability Theory to Domain-Referenced Testing, Iowa City : The American College Testing Program, Technical Bulletin, No. 30, 1978.

Capie, W., Johnson, C.E., Anderson, S.J., Ellett, C.D., and okey, J.R., Teacher Performance Assessment Instruments, Athens, GA : Teacher Assessment Project, University of Georgia, 1979.

Smith, P.L., Sampling Errors of Variance Component Estimates in Small Multifacet Generalizability Studies, Journal of Educational Statistics, 1978, 3, 319-346.

Smith, P.L., Some Approaches to Determining the Stability of Estimated Variance Components, A paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA, April 1980.

Rating	Scale Description
1.	Uses negative words or actions to discourage learners from giving responses or asking questions.
2.	Ignores learners who wish to be recognized or learner contributions are accepted without disagreement or further comment.
3.	Acknowledges learners who wish to be recognized and occasionally asks for learner responses or questions. Responses by the teacher are adequate.
4.	Asks for responses or questions frequently throughout the lesson and provided feedback to learners.
5.	In addition to the items in 4, the teacher incorporates learner responses and questions into activities.

Figure 1. Uses responses and questions from learners in teaching

Table 1

Estimates of Variance Components, Generalizability Coefficient and Brennan's Ø (λ) for Differing Sample Size : Competency One

Source of Variation	Sample Size								Mean	Standard Deviation	Coefficient of Variation
	250	350	450	500	550	600	650	700			
Teacher (T)	.165	.168	.168	.172	.175	.172	.164	.165	.169	.004	2.4
T x Indicator (I)	.020	.026	.026	.027	.026	.027	.027	.030	.026	.003	11.5
T x Rater (R)	.031	.027	.022	.020	.021	.026	.030	.031	.026	.005	19.2
T x Occasion (O)	.087	.076	.074	.073	.067	.066	.069	.069	.073	.007	9.6
T x I x R	.017	.021	.021	.019	.020	.019	.021	.022	.020	.002	10.0
T x R x O	.128	.134	.134	.136	.133	.131	.128	.125	.131	.004	4.0
T x I x O	.077	.066	.063	.059	.058	.061	.061	.061	.063	.006	9.5
Residual	.337	.330	.333	.332	.331	.326	.330	.327	.331	.003	1.0
Generalizability Coefficient	.62	.64	.65	.66	.67	.66	.65	.65	.65	.02	3.0
Brennan's Ø (λ)	.78	.78	.78	.79	.80	.79	.78	.79	.79	.01	1.3

Table 2

Estimates of Variance Components, Generalizability Coefficient, and Brennan's Ø (λ) for Replicated Samples of 50 and 150 : Competency One

Source of Variation	N = 50			N = 150		
	Mean	Standard Deviation	Coefficient of Variation	Mean	Standard Deviation	Coefficient of Variation
Teacher (T)	.185	.061	32.8	.165	.028	16.7
T x Indicator (I)	.022	.007	30.1	.027	.009	34.4
T x Rater (R)	.030	.021	69.8	.030	.021	69.7
T x Occasion (O)	.043	.034	78.6	.067	.015	22.8
T x I x R	.022	.018	82.1	.021	.008	40.5
T x R x O	.137	.044	32.3	.130	.011	8.6
T x I x O	.058	.021	35.4	.062	.007	11.4
Residual	.337	.027	8.1	.326	.021	6.5
Generalizability Coefficient	.69	.12	17.1	.65	.06	9.2
Brennan's Ø (λ)	.80	.06	8.1	.78	.03	3.9

Table 3

Estimates of Variance Components, Generalizability Coefficient, and Brennan's Ø (λ) for Differing Sample Size : Competency Two

Source of Variation	Sample Size									Mean	Standard Deviation	Coefficient of Variation
	200	250	300	350	400	450	500	600	700			
Teacher (T)	.057	.068	.065	.064	.068	.066	.067	.068	.073	.066	.004	6.1
T x Indicator (I)	.014	.014	.015	.018	.018	.020	.017	.019	.020	.017	.002	11.8
T x Rater (R)	.029	.029	.028	.025	.029	.031	.030	.028	.027	.028	.002	7.1
T x Occasion (O)	.022	.023	.028	.024	.023	.029	.027	.023	.022	.025	.003	12.0
T x I x R	.023	.030	.030	.030	.029	.027	.026	.023	.023	.026	.003	11.5
T x R x O	.063	.066	.060	.068	.070	.069	.070	.072	.073	.068	.004	5.9
T x I x O	.031	.030	.029	.029	.031	.025	.027	.025	.024	.028	.003	10.7
Residual	.274	.273	.269	.271	.273	.279	.280	.280	.281	.275	.004	1.45
Generalizability Coefficients	.54	.58	.57	.57	.57	.55	.56	.58	.60	.57	.02	3.5
Brennan's Ø (λ)	.53	.56	.54	.54	.54	.52	.53	.54	.56	.54	.01	3.7

Table 4

Estimates of Variance Components, Generalizability Coefficient, and Brennan's Ø (λ) for Replicated Samples of 50 and 150 : Competency Two

Source of Variation	N = 50			N = 150		
	Mean	Standard Deviation	Coefficient of Variation	Mean	Standard Deviation	Coefficient of Variation
Teacher (T)	.072	.039	53.7	.072	.015	21.0
T x Indicator (I)	.018	.013	71.1	.021	.006	27.5
T x Rater (R)	.032	.016	49.3	.029	.012	40.9
T x Occasion (O)	.028	.027	96.4	.022	.008	37.0
T x I x R	.020	.019	93.5	.021	.012	60.1
T x R X O	.065	.015	22.8	.069	.010	14.1
T x I x O	.030	.018	59.4	.021	.005	23.8
Residual	.287	.021	7.5	.284	.003	1.2
Generalizability Coefficient	.54	.17	31.8	.59	.06	9.7
Brennan's Ø (λ)	.53	.14	26.2	.56	.05	9.0